A hybrid-hierarchical genome assembly strategy to sequence the invasive golden mussel *Limnoperna fortunei*

Marcela Uliano-Silva¹²³*, Francesco Dondero⁴, Thomas Dan Otto⁵⁶, Igor Costa⁷, Nicholas Costa Barroso Lima⁷⁸, Juliana Alves Americo¹, Camila Junqueira Mazzoni²³, Francisco Prosdocimi⁷, Mauro de Freitas Rebelo¹*

Marcela Uliano-Silva: marcela.uliano@gmail.com
Francesco Dondero: francesco.dondero@uniupo.it
Thomas D. Otto: tdo@sanger.ac.uk
Igor Costa: igor.bioinfo@gmail.com
Nicholas Costa Barroso Lima: ncblima@gmail.com
Juliana Alves Americo: juliana.americo@gmail.com
Camila Mazzoni: mazzoni@izw-berlin.de
Francisco Prosdocimi: prosdocimi@bioqmed.ufrj.br
Mauro de Freitas Rebelo: mrebelo@biof.ufrj.br

Affiliations:

1 Carlos Chagas Filho Biophysics Institute (IBCCF), Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

2 Department of Evolutionary Genetics, Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany

3 Berlin Center for Genomics in Biodiversity Research, Berlin, Germany

4 Department of Science and Technological Innovation (DiSIT), Università del Piemonte Orientale Amedeo Avogadro, Vercelli-Novara-Alessandria, Italy
ABSTRACT

Background: For more than 25 years, the golden mussel *Limnoperna fortunei* has aggressively invaded South American freshwaters, having travelled more than 5,000 km upstream across five countries. Along the way, the golden mussel has outcompeted native species and economically harmed aquaculture, hydroelectric powers, and ship transit. We have sequenced the complete genome of the golden mussel to understand the molecular basis of its invasiveness and search for ways to control it.

Findings: We assembled the 1.6 Gb genome into 20548 scaffolds with an N50 length of 312 Kb using a hybrid and hierarchical assembly strategy from short and long DNA reads and transcriptomes. A total of 60717 coding genes were inferred from a customized transcriptome-trained AUGUSTUS run. We also compared predicted protein sets with those of complete molluscan genomes, revealing an exacerbation of protein-binding domains in *L. fortunei*.

Conclusions: We built one of the best bivalve genome assemblies available using a cost-

effective approach using Illumina pair-end, mate pair, and PacBio long reads. We expect that the continuous and careful annotation of *L. fortunei*’s genome will contribute to the investigation of bivalve genetics, evolution, and invasiveness, as well as to the development of biotechnological tools for aquatic pest control.

**KEYWORDS:** Amazon; binding domain; bivalves; genomics; TLR; transposon.

**DATA DESCRIPTION**

The golden mussel *Limnoperna fortunei* is an Asian bivalve that arrived in the southern part of South America about 25 years ago [1]. Research suggests that *L. fortunei* was introduced in South America through ballast water of ships coming from Hong Kong or Korea [2]. It was found for the first time in the estuary of the La Plata River in 1991 [1]. Since then, it has moved ~5,000 km, invading upstream continental waters and reaching northern parts of the continent [3] leaving behind a track of great economic impact and environmental degradation [4]. The latest infestation was reported in 2016 in the São Francisco River, one of the main rivers in the Northeast of Brazil, with a 2,700 km riverbed that provides water to more than 14 million people. At Paulo Afonso, one of the main hydroelectric power plants in the São Francisco River, maintenance due to clogging of pipelines and corrosion caused by the golden mussel is estimated to cost US$ 700,000 per year (*personal communication, Mizaël Gusmã, Chief Maintenance Engineer for Centrais Hidrelétricas do São Francisco – CHESF)*.

A recent review has shown that, before arriving in South America, *L. fortunei* was already an invader in China. Originally from the Pearl River Basin, the golden mussel has traveled 1,500 km into the Yang Tse and the Yellow River basins, being limited further north
only by the extreme natural barriers of Northern China [5]. Today, *L. fortunei* is found in the Paraguaizinho River, located only 150 km from the Teles-Pires River that belongs to the Alto Tapajós River Basin and is the first to directly connect with the Amazon River Basin [6]. Due to its fast dispersion rates, it is very likely that *L. fortunei* will reach the Amazon River Basin in the near future.

The reason why some freshwater bivalves, such as *L. fortunei*, *Dreissena polymorpha*, and *Corbicula fluminea*, are aggressive invaders is not fully understood. These bivalves present characteristics such as (i) tolerance to a wide range of environmental variables, (ii) short life span, (iii) early sexual maturation, and (iv) high reproductive rates that allow them to reach densities as high as 150,000 ind.m\(^{-2}\) over a year [7, 8] that may explain the aggressive behavior. On the other hand, these traits are not exclusive to invasive freshwater bivalves and do not explain how they outcompete native species and disperse so widely.

To the best of our knowledge, there are no reports of successful strategies to control the expansion of mussel invasion in industrial facilities. Bivalves can sense chemicals in the water and close their valves as a defensive response [9], making them tolerant to a wide range of chemical substances, including strong oxidants like chlorine [10]. Microencapsulated chemicals have shown better results in controlling mussel populations in closed environments [10, 11] but it is unlikely they would work in the wild. Currently, there is no effective and efficient approach to control the invasion by *L. fortunei*.

The genome sequence is one of the most relevant and informative descriptions of species biology. The genetic substrate of invasive populations, upon which natural selection operates, can be of primary importance to understand and control a biological invader [12, 13].
We have partially funded the golden mussel genome sequencing through a pioneer crowdfunding initiative in Brazil (www.catarse.me/genoma). In this campaign, we could raise around USD$ 20,000.00 at the same time we promoted scientific education and awareness in Brazil.

Here we present the first complete genome dataset for the invasive bivalve *Limnoperna fortunei*, assembled from short and long DNA reads and using a hybrid and hierarchical assembly strategy. This high-quality reference genome represents a substantial resource for further studies of genetics and evolution of mussels, as well as for the development of new tools for plague control.

**Genome sequencing in short Illumina and long PacBio reads**

*Limnoperna fortunei* mussels were collected from the Jacui River, Porto Alegre, Rio Grande do Sul, Brazil (29°59′29.3″S 51°16′24.0″W). Voucher specimens were housed at the zoological collection (specimen number: 19643) of the Biology Institute at the Universidade Federal do Rio de Janeiro, Brazil. For the genome assembly, a total of 3 individuals were sampled for DNA extraction from gills and to produce the three types of DNA libraries used in this study. DNA was extracted using DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) to prepare libraries for Illumina Nextera paired-end reads, with ~180bp and ~500bp of insert size, (ii) Illumina Nextera mate-pair reads with insert sizes from 3 to 15 Kb, and (iii) Pacific Biosciences long reads (Table 1). Illumina libraries were sequenced respectively in a HiScanSQ or HiSeq 1500 machine, and Pacific Biosciences reads were produced with the P4C6 chemistry and sequenced in 10 SMRT Cells. All Illumina reads were submitted to quality analysis with FastQC (FastQC, RRID:SCR_014583) followed by trimming with Trimmomatic (Trimmomatic,
Pacific Biosciences adaptor-free subreads sequences were used as input data for the genome assembly.

Table 1 - DNA reads produced for *L. fortunei* genome assembly

<table>
<thead>
<tr>
<th>Library technology</th>
<th>Reads insert size</th>
<th>Pairs</th>
<th>Raw data</th>
<th>Trimmed Data*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of reads</td>
<td>Number of bases</td>
<td>Number of reads</td>
<td>Number of bases</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Illumina Nextera</td>
<td>Paired end – 180 bp</td>
<td>R1</td>
<td>209542721</td>
<td>21060365702</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R2</td>
<td>209542721</td>
<td>21049308698</td>
</tr>
<tr>
<td></td>
<td>Paired end – 500 bp</td>
<td>R1</td>
<td>153948902</td>
<td>15472966961</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R2</td>
<td>153948902</td>
<td>15462883157</td>
</tr>
<tr>
<td></td>
<td>Mate pair 3-12 Kb</td>
<td>R1</td>
<td>178392944</td>
<td>18017687344</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R2</td>
<td>178392944</td>
<td>18017687344</td>
</tr>
<tr>
<td>Pacific Biosciences</td>
<td>P4C - 10/SMTRC</td>
<td>Subreads</td>
<td>1663730</td>
<td>11171487485</td>
</tr>
</tbody>
</table>


For transcriptome sequencing, RNA was sampled from four tissues (gills, adductor muscle, digestive gland, and foot) of three different golden mussel specimens. RNA was
extracted using NEXTflex Rapid Directional RNA-Seq Kit (Bioo Scientifics, TX, USA) and 12 barcodes from NEXTflex Barcodes compatible with Illumina NexSeq Machine. Resulting reads (Supplementary Table S1) were submitted to FastQC quality analysis and trimmed with Trimmomatic for all NEXTflex adaptors and barcodes. A total of 3 sets of de novo assembled transcriptomes were generated using Trinity (Trinity, RRID:SCR_013048) (Table 2); one set for each specimen was a pool of the 4 tissue samples to avoid assembly bias due to intraspecific polymorphism [15].

Table 2 - Trinity assembled transcripts used in the assembly and annotation of L. fortunei genome

<table>
<thead>
<tr>
<th>Sample</th>
<th>Pooled tissues</th>
<th>Number of reads prior assembly</th>
<th>Number of Trinity Transcripts</th>
<th>Number of Trinity Genes</th>
<th>Average Contig Length</th>
<th>GC%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mussel 1</td>
<td>Gills, mantle, digestive gland, foot</td>
<td>406589144</td>
<td>433197</td>
<td>303172</td>
<td>854</td>
<td>34</td>
</tr>
<tr>
<td>Mussel 2</td>
<td>Gills, mantle, digestive gland, foot</td>
<td>376577660</td>
<td>435054</td>
<td>298117</td>
<td>824</td>
<td>34</td>
</tr>
<tr>
<td>Mussel 3</td>
<td>Gills, mantle, digestive gland, foot</td>
<td>334316116</td>
<td>499392</td>
<td>351649</td>
<td>844</td>
<td>34</td>
</tr>
</tbody>
</table>

Genome assembly using a hybrid and hierarchical strategy
The Jellyfish software (Jellyfish, RRID:SCR_005491) [16] was used to count and determine the distribution frequency of lengths 25 and 31 k-mers (Figure 1) for the Illumina DNA paired-end and mate-pair reads (Table 1). Genome size was estimated to be 1.6 Gb by using the 25 k-mer distribution plot as total k-mer number and then subtracting erroneous reads (starting k-mer counts from 12 times coverage), to further divide by the homozygous coverage-peak depth (45 times coverage), as performed by Li et al. (2010) [17]. A double-peak k-mer distribution was used as evidence of genome diploidy (Figure 1) and high heterozygosity. The rate of heterozygosity was estimated to be 2.3% and it was calculated as described by Vij et al. (2016) [18], using as input data the 17-kmer distribution plot for reads from one unique specimen.

Initially, we attempted to assemble the golden mussel genome using only short Illumina reads of different insert sizes (paired-end and mate-pairs, Table 1) using traditional de novo assembly software such as ALLPATHS (ALLPATHS-LG, RRID:SCR_010742) [19], SOAPdenovo (SOAPdenovo, RRID:SCR_010752) [20], and MaSuRCA (MaSuRCA, RRID:SCR_010691) [21]. All these attempts resulted in very fragmented genome drafts, with an N50 no higher than 5 Kb and a total of 4 million scaffolds. To reduce fragmentation, we further sequenced additional long reads (10 PacBio SMTR Cells, Table 1) and performed a hybrid and hierarchical de novo assembly described below and depicted in Figure 2.

First, (i) trimmed paired-end and mate-pair DNA Illumina reads (Table 1) were assembled into contigs using the software Sparse Assembler [22] with parameters LD 0 NodeCovTh 1 EdgeCovTh 0 k 31 g 15 PathCovTh 100 GS 1800000000. Next, (ii) the resulting contigs were assembled into scaffolds using Pacific Biosciences long subreads data and the PacBio-correction-free assembly algorithm DBG2OLC [23] with parameters LD1 0 k 17.
KmerCovTh 10 MinOverlap 20 AdaptiveTh 0.01. Finally, (iii) resulting scaffolds were submitted to 6 iterative runs of the program L_RNA_Scaffolder [24] that uses exon-distance information from de novo assembled transcripts (Table 2) to fill gaps and connect scaffolds whenever appropriate. At the end, (iv) the final genome scaffolds were corrected for Illumina and Pacific Biosciences sequencing errors with the software PILON [25]: all DNA and RNA short Illumina reads were re-aligned back to the genome with BWA aligner (BWA, RRID:SCR_010910) [26] and resulting sam files were BAM-converted, sorted, and indexed with samtools package (SAMTOOLS, RRID:SCR_002105) [27]. Pilon [25] identifies INDELS and mismatches by coverage of reads and yields a final corrected genome draft. Pilon was run with parameters -- diploid -- duplicates.

The final genome was assembled in 20,548 scaffolds, with an N50 of 312 Kb and a total assembly length of 1.6 Gb (Table 3).

Table 3: Assembly statistics for Limnoperna fortunei’s genome

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated genome size by k-mer analysis</td>
<td>1.6 Gb</td>
</tr>
<tr>
<td>Total size of assembled genome</td>
<td>1.673 Gb</td>
</tr>
<tr>
<td>Number of scaffolds</td>
<td>20548</td>
</tr>
<tr>
<td>Number of contigs</td>
<td>61093</td>
</tr>
<tr>
<td>Scaffold N50</td>
<td>312 Kb</td>
</tr>
<tr>
<td>Maximum scaffold length</td>
<td>2.72 Mb</td>
</tr>
<tr>
<td>Percentage of genome in scaffolds &gt; 50 Kb</td>
<td>82.55%</td>
</tr>
</tbody>
</table>
The golden mussel genome presents 81% of all Benchmarking Universal Single Copy Orthologs (BUSCO version 3.3 analysis with Metazoa database; BUSCO, RRID:SCR_015008) (Table 4) and, compared to the mollusk genomes currently available [28, 29, 30, 31, 32, 33, 34, 35] it represents one of the best assemblies of molluscan genomes so far also in terms of scaffold N50 and contiguity (Table 5).

One main challenges of assembling bivalve genomes lies in the high heterozygosity and amount of repetitive elements these organisms present: (i) the mussels *L. fortunei* and *Modiolus philippinarum* and the oyster *Crassostrea gigas* genomes were estimated to have heterozygosity rates of 2.3%, 2.02% 1.95% respectively, which is substantially higher than other animal genomes [29], and (ii) repetitive elements correspond to at least 30% of the genomes of all studied bivalves so far (Table 3) [28, 29, 30, 31, 33, 34, 35]. Also, retroelements might be active in some species such as *L. fortunei* (refer to the retroelements-related section of this paper) and *C. gigas* [29], allowing genome rearrangements that may hinder for genome assembly. One exception seems to be the deep-sea mussel *B. platifrons* which has lower heterozygosity rates compared to other bivalves [31]. Sun *et al.*, (2017) [31] suggested it might be due to recurrent population bottlenecks happened after events of population extinction and recolonization in the extreme environment [31]. Nevertheless, most of the bivalve genome projects relying only on short Illumina reads are likely to present fragmented initial drafts [28, 30]. PacBio long reads allowed us to increase the N50 to 32 Kb and to reduce the number of scaffolds from millions to
61102, using the DBG2OLC [23] assembler. Finally, interactive runs of L_RNA_scaffolder [24] using the transcriptomes (Table 2) rendered the final result of N50 312 Kb in 20548 scaffolds. It’s important to note that assembly statistics can perform better for genomes assembled with reads generated with DNA extracted from one unique individual. This, however, was not possible for L. fortunei’s genome, due to the high amount of high-quality-DNA necessary to produce Illumina mate-pair and PacBio long reads. In this study, the challenge of assembling the high polymorphic regions between haplotypes was enhanced by the difficulties of assembling reads originated from highly polymorphic regions across individuals. However, the golden mussel assembly presented here shows that the use of Illumina contigs, low coverage of PacBio long reads, transcriptome and Illumina re-mapping for final correction (Figure 2) represents an option for cost-efficient assembly of highly heterozygous genomes of nonmodel species such as bivalves.

Table 4: Summary statistics of Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis for L. fortunei genome run for Metazoans

<table>
<thead>
<tr>
<th>Categories</th>
<th>Number of Genes</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total BUSCO groups searched</td>
<td>978</td>
<td>--</td>
</tr>
<tr>
<td>Complete BUSCOs</td>
<td>801</td>
<td>81.9%</td>
</tr>
<tr>
<td>Complete and single-copy BUSCOs</td>
<td>769</td>
<td>78.62%</td>
</tr>
<tr>
<td>Complete and duplicated BUSCOs</td>
<td>32</td>
<td>3.27%</td>
</tr>
<tr>
<td></td>
<td>Haliotis discus hannai</td>
<td>Lottia gigantea</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td><strong>Estimated genome size</strong></td>
<td>1.65 Gb</td>
<td>359.5 Mb</td>
</tr>
<tr>
<td><strong>Number of scaffolds</strong></td>
<td>80,032</td>
<td>4,475</td>
</tr>
<tr>
<td><strong>Total size of scaffolds</strong></td>
<td>1,865,475,499</td>
<td>359,512,207</td>
</tr>
<tr>
<td><strong>Longest scaffold</strong></td>
<td>2,207,537</td>
<td>9,386,848</td>
</tr>
<tr>
<td><strong>Shortest scaffold</strong></td>
<td>854</td>
<td>1,000</td>
</tr>
<tr>
<td><strong>Number of scaffolds &gt; 1 K nt</strong></td>
<td>79,923</td>
<td>4,471</td>
</tr>
<tr>
<td></td>
<td>(99.9%)</td>
<td>(99.9%)</td>
</tr>
<tr>
<td><strong>Number of scaffolds &gt; 1 M nt</strong></td>
<td>67</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>(0.1%)</td>
<td>(2.2%)</td>
</tr>
</tbody>
</table>

Table 5: Comparison of genome assembly statistics for molluscan genomes.
<table>
<thead>
<tr>
<th></th>
<th>Mean scaffold size</th>
<th>Median scaffold size</th>
<th>N50 scaffold length</th>
<th>Sequencing coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>23,309</td>
<td>1,697</td>
<td>200,099</td>
<td>322 X</td>
</tr>
<tr>
<td></td>
<td>80,338</td>
<td>3,622</td>
<td>1,870,055</td>
<td>8.87 X</td>
</tr>
<tr>
<td></td>
<td>81,655</td>
<td>13,763</td>
<td>264,327</td>
<td>11 X</td>
</tr>
<tr>
<td></td>
<td>11,441</td>
<td>1,327</td>
<td>48,447</td>
<td>39.7 X</td>
</tr>
<tr>
<td></td>
<td>11,939</td>
<td>362</td>
<td>803,631</td>
<td>297 X</td>
</tr>
<tr>
<td></td>
<td>46,671</td>
<td>824</td>
<td>401,319</td>
<td>155 X</td>
</tr>
<tr>
<td></td>
<td>114,508</td>
<td>14,683</td>
<td>345,846</td>
<td>234 X</td>
</tr>
<tr>
<td></td>
<td>916</td>
<td>258</td>
<td>2,651</td>
<td>32 X</td>
</tr>
<tr>
<td></td>
<td>25,269</td>
<td>1,284</td>
<td>343,373</td>
<td>319 X</td>
</tr>
<tr>
<td></td>
<td>35,262</td>
<td>13,722</td>
<td>100,161</td>
<td>209.5 X</td>
</tr>
<tr>
<td></td>
<td><strong>81,425</strong></td>
<td></td>
<td><strong>312,020</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Sequencing Technology</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Illumina + PacBio</td>
<td>Sanger</td>
<td>Sanger</td>
<td>Illumina + BACs</td>
</tr>
<tr>
<td></td>
<td><strong>Illumina</strong></td>
<td><strong>Illumina</strong></td>
<td><strong>Illumina</strong></td>
<td><strong>Illumina</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Illumina</strong></td>
<td><strong>Illumina</strong></td>
<td><strong>Illumina</strong></td>
<td><strong>Illumina</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Illumina + PacBio</strong></td>
<td><strong>Illumina</strong></td>
<td><strong>Illumina</strong></td>
<td><strong>Illumina</strong></td>
</tr>
</tbody>
</table>

**Around 10% of repetitive elements are transposons**

Initial masking of *L. fortunei* genome was done using RepeatMasker program (RepeatMasker, RRID:SCR_012954) [36] with parameter `-species bivalves` and masked 3.4% of the total genome. This content was much lower than the masked portion of other molluscan genomes: 34% in *C. gigas* [29] and 36% in *M. galloprovincialis* [28], suggesting that the fast evolution of interspersed elements limits the use of repeat libraries from divergent taxa [37]. Thus, we generated a *de novo* repeat library for *L. fortunei* using the program RepeatModeler (RepeatModeler, RRID:SCR_015027) [38] and its integrated tools (RECON [39], TRF [40], and RepeatScout [41]). This *de novo* repeat library was the input to RepeatMasker together with the first masked genome draft of *L. fortunei*, and resulted in a final masking of 33.4% of the genome. Even though more than 90% of the repeats were not classified by RepeatMasker (Supplementary Table S2), 8.85% of the repeats were classified as LINEs, Class I transposable elements. In addition, large numbers of reverse-transcriptases (824 counts, Pfam RVT_1 PF00078), transposases (177 counts, Pfam HTH_Tnp_Tc3_2 PF01498), and integrases (501
counts, Pfam Retroviral integrase core domain PF00665) and other related elements were detected; over 98% of these had detectable transcripts.

More than 30,000 sequences identified by gene prediction and automated annotation.

To annotate the golden mussel genome, we sequenced a number of transcriptomes (Table S1), de novo assembled (Table 2) and aligned these transcriptomes to the genome scaffolds, and created gene models with the PASA pipeline [36]. These models were used to train and run the ab initio gene predictor AUGUSTUS (Augustus: Gene Prediction, RRID:SCR_008417) [37] (Supplementary Figure S1). The complete gene models yielded by PASA [42] were BLASTed (e-value 1e-20) against the Uniprot database (UniProt, RRID:SCR_002380) and those with 90% or more of their sequences showing in the BLAST hit alignment were considered for further analysis. Next, all the necessary filters to run an AUGUSTUS [43] personalized training were performed: (i) only gene models with more than 3 exons were maintained, (ii) sequences with 90% or more overlap were withdrawn and only the longest sequences were retained, and (iii) only gene models free of repeat regions, as indicated by BLASTN similarity searches with de novo library of repeats, were maintained. These curated data yielded a final set of 1,721 gene models on which AUGUSTUS [35] was trained in order to predict genes in the genome using the default AUGUSTUS [43] parameters. Once the gene models were predicted, a final step was performed by using the PASA pipeline [42] once again in the update mode (parameters -c -A -g -t). This final step compared the 55,638 gene models predicted by AUGUSTUS [43] with the 40,780 initial transcript-based gene-structure models from PASA [42] to generate the final set of
60,717 gene models for *L. fortunei*. Of those, 58% had transcriptional evidence based on RNA Illumina reads (Table S2) re-mapping, rate that was expected since our RNA-Seq libraries were constructed only for 4 tissues of adult golden mussel specimens without any environmental stresses induction (Table 2). Therefore, these libraries lack transcripts for developmental stages, for some other cell types (i.e. hemocytes) and stress-inducible genes. Finally, 67% of the gene models were annotated by homology searches against Uniprot or NCBI NR (Table 6).

Table 6: Summary of gene annotation against various databases for *L. fortunei* whole genome-predicted genes

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of genes</td>
<td>60,717</td>
</tr>
<tr>
<td>Total number of exons</td>
<td>220,058</td>
</tr>
<tr>
<td>Total number of proteins</td>
<td>60,717</td>
</tr>
<tr>
<td>Average protein size</td>
<td>304 aa</td>
</tr>
<tr>
<td>Number of protein BLAST hits* with Uniprot</td>
<td>26,198</td>
</tr>
<tr>
<td>Number of protein BLAST hits* with NR NCBI (no hits with Uniprot)</td>
<td>14,810</td>
</tr>
<tr>
<td>Number of protein HMMER hits* with Pfam.A</td>
<td>24,513</td>
</tr>
<tr>
<td>Number with proteins with KO assigned by KEGG</td>
<td>8,387</td>
</tr>
<tr>
<td>Number of proteins with BLAST hits* with EggNOG</td>
<td>36,868</td>
</tr>
</tbody>
</table>
*all considered hits had a minimum e-value of 1e-05

Protein clustering indicates evolutionary proximity among mollusks species.

Gene family relationships were assigned using reciprocal best BLAST and OrthoMCL software (version 1.4) [44] between L. fortunei proteins and the total protein set predicted for nine other mollusks: the mussels M. galloprovincialis, M. philippinarum and B. platifrons, the clam Ruditapes philippinarum, the scallop Patinopecten yessoensis, the pacific oyster C. gigas, the pearl oyster Pinctada fucata (genome version from Du et al [35]), and the gastropods Lottia gigantea and Haliotis discus hannai (see Supplementary Table S3 for detailed information on the comparative data). Figure 3A presents orthologs relationships for five of the bivalves analyzed. A total of 6,337 orthologs groups are shared among the five bivalve species.

Of all the orthologous found for the total 10 species, 44 groups are composed of single-copy orthologs containing one representative protein sequence of each species. These sequences were used to reconstruct a phylogeny: the single-copy orthologs sequences were concatenated and aligned with CLUSTALW [45] with a resulting alignment of 30755 sites in length (Figure 3B). ProtTest 3.4.2 [46] was used to estimate the best fitting substitution model, which was VT+G+I+F [47]. With this alignment and model we reconstructed the phylogeny using PhyML [48] and 100 bootstrap repetition, the resulting tree is shown on Figure 3B.

Protein domain analysis shows expansion of binding domain in L. fortunei.
We performed a quantitative comparison of protein domains predicted from whole genome projects of 10 molluscan species. The complete protein sets of *M. galloprovincialis*, *M. philippinarum* and *B. platifrons*, *Ruditapes philippinarum*, *P. yessoensis*, *C. gigas*, *Pinctada fucata*, *Lottia gigantea* and *Haliotis discus hannai* (Supplementary Table S3) were submitted to domain annotation using HMMER against Pfam-A database (e-value 1e-05). Protein expansions in *L. fortunei* were rendered using the normalized Pfam count value (average) obtained from the other nine mollusks, according to a model based on the Poisson cumulative distribution. Bonferroni correction (*p ≤ 0.05*) was applied for false discovery and absolute frequencies of Pfam-assigned-domains were initially normalized by the total count number of Pfam-assigned-domains found in *L. fortunei* to compensate for discrepancies in genome size and annotation bias.

For *L. fortunei*, the annotation against Pfam.A classified 40127 domains in 24513 gene models of which 83 and 67 were respectively expanded or contracted in comparison with the other mollusks (Supplementary Table S4 and S5; Figure 4A). The 83 overrepresented domains were further analyzed for functional enrichment using domain-centric Gene Ontology (Figure 4B). The analysis shows a prominent expansion of binding domains in *L. fortunei*, such as Thrombospondin (TSP_1), Collagen, Immunoglobulins (Ig, I-set, Izumo-Ig Ig_3), and Ankyrins (Ank_2, Ank_3, and Ank_4). These repeats have a variety of binding properties and are involved in cell-cell, protein-protein and receptor-ligand interactions driving evolutionary improvement of complex tissues and immune defense system in metazoans [49, 50, 51, 52, 53]. An evolutionary pressure towards the development of a diversified innate immune system is also suggested by the high amount of Leucine Rich Repeats (LRR) and Toll/interleukin-1 receptor homology domains (TIR). Death, another over-represented PFAM, is also part of TLR signaling, being
present in several docking proteins such as Myd88, Irak4 and Pelle [54]. Interestingly, BLAST analysis of *L. fortunei* gene models against Uniprot identified two types of Toll Like Receptors (TLRs) whose prototypical architecture of N-terminal extracellular leucine-rich repeat (LRR) motifs including either a single or multiple cysteine cluster domain, a C-terminal TIR domain spaced by a single transmembrane-spanning domain [55] could be correctly identified using the Simple Modular Architecture Research Tool (SMART) [56]. Indeed, we confirmed 141 sequences with similarity to single cysteine clusters TLRs (scc) typical of vertebrates, and 29 sequence hits with the multiple cysteine cluster TLRs (mcc) typical of *Drosophila* [55].

Phylogenetic analysis of all sequences (using PhyML [48], model JTT) (Supplementary Figure S2) shows evidence for TLRs clade separation in *L. fortunei*; the scc TLRs exhibit a higher degree of amino acid changes, higher molecular evolution, and diversification than the mcc TLRs. Overall, the expansion of these gene families might suggest an improved resistance to infections. It is, however, equally curious that other immune-related gene families such as Fribinogen_C and C1q seem to be contracted (Supplementary Table S5). This feature may depend on the evolutionary-driven, yet random, fate of the *L. fortunei* genome and consequence of different specific duplicate genes in other species. Also, other protein families involved in toxin metabolism, especially glutathione based processes and sulfotransferases are clearly contracted (Table S5).

**Final considerations**

Here we have described the first version of the golden mussel complete genome and its automated gene prediction that were funded through a crowdfunding initiative in Brazil. This genome contains valuable information for further evolutionary studies of bivalves and metazoa in general. Additionally, our team will further search for the presence of proteins of
biotechnology interest such as the adhesive proteins produced by the foot gland that we have
described elsewhere [57], or genes related to the reproductive system that have been shown to be
very effective for invertebrate plague control [58]. The golden mussel genome and the predicted
proteins are available for download in the Gigabase repository and the scientific community is
welcome to further curate the gene predictions.

As the golden mussel advances towards the Amazon river basin, the information provided in this
study may be used to help developing biotechnological strategies that may control the expansion
of this organism in both industrial facilities and open environment.

Availability of supporting data

*Limnopena fortunei’s* genome and transcriptome data are available in the Sequence
Read Archive (SRA) as BioProject PRJNA330677 and under the accession numbers
**SRR5188384, SRR5195098, SRR518800, SRR5195097, SRR5188315, SRR5181514**. This
Whole Genome Shotgun project has been deposited in the DDBJ/ENA/GenBank under accession
number NFUK00000000. The version described in this paper is version NFUK01000000.

Supporting data, also including annotations and BUSCO results, are available via the
*GigaScience* repository GigaDB [59].

Additional files

**Supplementary Table S1.** RNA raw reads sequenced for 3 *L. fortunei* specimens, 4 tissues each.
Supplementary Table S2: RepeatMasker classification of repeats predicted in *L. fortunei* genome.

Supplementary Table S3: Details of the online availability of the data used for ortholog assignment and protein domain expansion analysis.

Supplementary Table S4: Expanded protein families in *L. fortunei* genome.

Supplementary Table S5: Contracted protein families in *L. fortunei* genome.

Supplementary Table S6: Fantasy names given to *L. fortunei* genes and proteins from the backers that have supported us through crowdfunding (www.catarse.me/genoma).

Supplementary Figure 1: Steps performed for the prediction and annotation of *L. fortunei* genome.

Supplementary Figure 2: Phylogenetic tree of Toll-like (TLRs) receptors found in *L. fortunei* genome.

List of Abbreviations

BUSCO: Benchmarking Universal Single-Copy Orthologs; SRA: Sequence Read Archive; KEGG: Kyoto Encyclopedia of Genes and Genomes.

Competing interests

The authors declare that they have no competing interests.

Authors’ contribution
Conceived and designed the experiments: MR, MU, TO, CM, FD. Performed the experiments: MU, JA. Analyzed the data: MU, TO, CM, FD, FP, NC, IC, MR. Contributed reagents/materials/analysis tools: MR, FP, CM. Wrote the paper: MU, FD, MR. All authors read and approved the final manuscript.

**Funding**

This work was supported by the Brazilian Government agencies CAPES (PVE 71/2013), FAPERJ APQ1 (2014), and FAPERJ/DFG (39/2014). Also, this work was funded through crowdfunding with the support of 346 people ([www.catarse.me/genoma](http://www.catarse.me/genoma)).

**Acknowledgements**

We thank Susan Mbedi and Kirsten Richter from BeGenDiv for RNA-Seq library preparation and sequencing. We thank Dr. Loris Bennett for IT support while performing bioinformatics analysis.

We especially want to thank the 346 backers that supported the sequencing of the golden mussel through crowdfunding, in a 2013 campaign that raised U$ 20,000.00 ([www.catarse.me/genoma](http://www.catarse.me/genoma)).

We decided to give fantasy names to the genes and proteins that we found in the genome, to thank the backers for their support. The name list is available in **Supplementary Table S6**.

**Consent for publication**

Does not apply.
Ethics approval

*Limnoperna fortunei* specimens used for DNA extraction and sequencing were collected in the Jacuí River (29°59′29.3″S 51°16′24.0″W), southern Brazil. This bivalve is an exotic species in Brazil and is not characterized as an endangered or protected species.

References


Figure 1: K-mer distribution of *Limnoperna fortunei* Illumina DNA reads (Table 1).
Figure 2: Hierarchical assembly strategy employed for the golden mussel genome assembly. Trimmed Illumina reads were assembled to the level of contigs with Sparse Assembler algorithm (Step 1). Then, Illumina contigs and PacBio reads were used to build scaffolds with DBG2OLC assembler, that anchors Illumina contigs to erroneous PacBio subreads, correcting them and building longer scaffolds (Step 2), followed by transcriptome joining scaffolds using L_RNA_scaffolder (Step 3). Final scaffolds were corrected by re-aligning all Illumina DNA and RNA-seq reads back to them and calling consensus with Pilon software (Step 4). In bold is bioinformatics software used in each step. Red blocks indicate PacBio errors, which are represented by insertions and/or deletions found in approximately 12% of PacBio subreads.
**Figure 3A:** Gene family assigned with OrthoMCL for the total set of proteins predicted from five mussel genome projects. Outside the Venn diagram its represented the species name and below it is the number of proteins / number of clustered proteins / number of clusters. **B:** Phylogeny of the concatenated data set using 44 single-copy orthologs extracted from ten molluscan genomes. The VT model was estimated to be best fitting substitution model with ProtTest 3.4.2. We reconstructed the phylogeny using PhyML and 100 bootstrap repetition.
Figure 4: Gene family representation analysis in the *L. fortunei* genome. Panel A. **PFAM hierarchical clustering, heatmap.** Features were selected according to a model based on the Poisson cumulative distribution of each PFAM count in the golden mussel genome vs the normalized average values found in the other nine molluscan genomes (Bonferroni correction, P ≤ 0.05). Transposable elements were included in the analysis. Colors depict the log2 ratio between PFAM counts found in each single genome and the corresponding mean value. The hierarchical clustering used the average dot product for data matrix and complete linkage for branching. Legend: Lf, *L. fortunei*; Bp, *Bathymodioulus platifrons*; Mg, *Mytilus galloprovincialis*; Mp, *Modioulus philippinarum*; Cg, *Crassostrea gigas*; Pf, *Pinctada fucata*; Py, *Patinopecten yessoensis*; Rp, *Ruditapes philippinarum*; Hd, *Haliotus discus hannai*; Lg,
*Lottia gigantea* Panel B. Gene ontology analysis of expanded gene families (PFAMs), **semantic scatter plot.** Shown are cluster representatives after redundancy reduction in a two-dimensional space applying multidimensional scaling to a matrix of semantic similarities of GO term. Color indicates the GO enrichment level (legend in upper left-hand corner); size indicates the relative frequency of each term in the UNIPROT database (larger bubbles represent less specific processes).